

Comparison of ChatGPT-3.5, ChatGPT-4, and Orthopaedic Resident Performance on Orthopaedic Assessment Examinations

Year: 2023 | Citations: 122 | Authors: P. Massey, Carver Montgomery, Andrew S Zhang

Abstract

Introduction: Artificial intelligence (AI) programs have the ability to answer complex queries including medical profession examination questions. The purpose of this study was to compare the performance of orthopaedic residents (ortho residents) against Chat Generative Pretrained Transformer (ChatGPT)-3.5 and GPT-4 on orthopaedic assessment examinations. A secondary objective was to perform a subgroup analysis comparing the performance of each group on questions that included image interpretation versus text-only questions.

Methods: The ResStudy orthopaedic examination question bank was used as the primary source of questions. One hundred eighty questions and answer choices from nine different orthopaedic subspecialties were directly input into ChatGPT-3.5 and then GPT-4. ChatGPT did not have consistently available image interpretation, so no images were directly provided to either AI format. Answers were recorded as correct versus incorrect by the chatbot, and resident performance was recorded based on user data provided by ResStudy.

Results: Overall, ChatGPT-3.5, GPT-4, and ortho residents scored 29.4%, 47.2%, and 74.2%, respectively. There was a difference among the three groups in testing success, with ortho residents scoring higher than ChatGPT-3.5 and GPT-4 ($P < 0.001$ and $P < 0.001$). GPT-4 scored higher than ChatGPT-3.5 ($P = 0.002$). A subgroup analysis was performed by dividing questions into question stems without images and question stems with images. ChatGPT-3.5 was more correct (37.8% vs. 22.4%, respectively, OR = 2.1, $P = 0.033$) and ChatGPT-4 was also more correct (61.0% vs. 35.7%, OR = 2.8, $P < 0.001$), when comparing text-only questions versus questions with images. Residents were 72.6% versus 75.5% correct with text-only questions versus questions with images, with no significant difference ($P = 0.302$).

Conclusion: Orthopaedic residents were able to answer more questions accurately than ChatGPT-3.5 and GPT-4 on orthopaedic assessment examinations. GPT-4 is superior to ChatGPT-3.5 for answering orthopaedic resident assessment examination questions. Both ChatGPT-3.5 and GPT-4 performed better on text-only questions than questions with images. It is unlikely that GPT-4 or ChatGPT-3.5 would pass the American Board of Orthopaedic Surgery written examination.